Linear regression using both temporally aggregated and temporally disaggregated data: Revisited

Qian, Hang
Iowa State University

July 2010

Linear Regression Using Both Temporally Aggregated and Temporally Disaggregated Data: Revisited

Hang Qian

Abstract

This paper discusses regression models with aggregated covariate data. Reparameterized likelihood function is found to be separable when one endogenous variable corresponds to one instrument. In that case, the full-information maximum likelihood estimator has an analytic form, and thus outperforms the conventional imputed value two-step estimator in terms of both efficiency and computability. We also propose a competing Bayesian approach implemented by the Gibbs sampler, which is advantageous in more flexible settings where the likelihood does not have the separability property.

Keywords: Aggregated covariate, Maximum likelihood, Bayesian inference

1. Introduction

Incomplete data is a common problem in applied economics. In the regression analysis, there are occasions in which complete data of many relevant regressors are collected, but data on one or more key covariates are aggregated by household, or by group, by region, by time, and so on. To make the best use of the available data and minimize information loss, we hope to use both aggregated and disaggregated data in a regression. The conventional wisdom is a two-step regression in which the first regression imputes the ag-

gregated data, and then the imputed covariate data are used in the second regression. Three decades ago, with the same title, Hsiao (1979) and Palm and Nijman (1982) considered the maximum likelihood (ML) estimation of an aggregated covariate data (ACD) model in which data are measured at different temporal frequencies. However, this approach received little attention in the subsequent empirical work. The least squares (LS) imputation remains the typical solution. Perhaps part of the reason is that Palm and Nijman found that the likelihood function cannot be factorized into two parts as suggested by Hsiao. So Palm and Nijman concluded that the computational advantage of ML is lost in the ACD model.

This paper revisits the model in Hsiao (1979). One contribution of this paper is that the likelihood function is found to be separable by suitable reparameterization if one instrument corresponds to one endogenous regressor. In that case, an analytic full-information ML estimator does exist. That implies the efficient estimator can be obtained without computational barriers, and thus overshadows the least squares imputation approach.

Our idea of likelihood separability is borrowed from the statistics literature addressing missing data. In contrast to the standard missing data problem where a fraction of observations are unavailable, data aggregation eliminates all the individual-level observations, leaving relatively few aggregated data. However, once the correlation structure of aggregated and disaggregated values is accounted for, the ACD regression bears many similarities to the missing data problem. Anderson (1957), in the context of missing multivariate normal variates, raised the important idea of factoring the likelihood function into two parts, each of which can be maximized analytically.

Gourieroux and Monfort (1981) extended that method to regression models with missing covariate data. In this paper, we follow this track and extend the idea of likelihood separability to the ACD model.

On top of that, we are aware that not every ACD model specification satisfies the likelihood separability conditions. Furthermore, the practitioners may have their own models while at the same time an aggregated covariate is involved. In that case, the likelihood function has to be maximized numerically. As an alternative to numerical ML, we propose a competing Bayesian approach implemented by the Gibbs sampler, which is another contribution of this paper. For models without analytic solutions, our Monte Carlo study shows that the Bayesian estimator is more robust and less sensitive to the initial values.

Our third contribution is a critique on LS imputation approaches applied to the ACD model. The asymptotics of LS-type estimators have been extensively discussed in the literature. Gourieroux and Monfort (1981) provided asymptotic comparisons of a variety of LS estimators for the missing data regression. In Hsiao (1979) and Palm and Nijman (1982) there are also comparisons the relative efficiency of ML estimator with LS-type estimators for the ACD regression. In addition, simulation-based multiple imputation strategies (see Rubin, 1987; Schafer, 1997; Allison, 2000) can also be used to compute the asymptotic standard error of those estimators. However, for a ACD regression model with endogenity problems, some LS-type estimators are not consistent, and some consistent estimators discard apparent information. Those drawbacks are overcome by the ML and Bayesian estimators, which is the main reason we do not recommend the usage of LS estimators.

The model in Hsiao (1979) is originally designed for temporal aggregation, which is commonly found in macroeconomic and financial data. Temporal aggregation and mixed sampling frequencies regression can be appropriately tackled by time series techniques. Geweke (1978), Ghysels et al. (2006) and Andreou et al. (2010) developed corresponding models and estimation techniques. Revisiting Hsiao (1979), we feel that his model might be most suitable for aggregation problems encountered in applied microeconomics. We illustrate the potential applications of the ACD regression with three examples.

Example 1. *We want to evaluate the impact of the Low-Income Home Energy Assistance Program (LIHEAP) on the subsequent energy expenditures of its recipients. The LIHEAP grant is a one-time payment for the winter season, whereas the gas or electricity is always billed monthly. Although we can aggregate the monthly bills as well and conduct an analysis at the seasonal level, we lose the monthly information contained in the dependent variable,and other covariates such as monthly income and weather. Now consider the ACD regression: monthly usage of the grant (in consumption or saving) is latent, up to the individual choice and summing up to the observable total amount. If we can impute the latent monthly grant usage, we will know what proportion of the grant contributes to monthly energy expenditures.*

Example 2. *Occupational Outlook Handbook (Bureau of Labor Statistics, U.S. Department of Labor, 2010) predicts that veterinarians will increase by 33% over the 2008–18 decade, much faster than the average for all occupations. Suppose we want to study whether the fast growth of the cat (pet) population pushes up the demand for veterinary services. We searched the database for public use and found that the veterinarian data, along with many*

other covariates, of each county are available, while the pet population is only recorded for each state, hence the ACD.

Example 3. *In development economics, we might be interested in the calorie-income elasticity in poor countries. The calorie intake is an individual measure, varying among men, women and children. However, the observable household income is likely to be redistributed within the family, and thus the real individual income is a latent regressor to the researcher.*

There are many practical reasons the covariate data are aggregated. In Example 1 and 3, by the nature of the variable, the disaggregated values are never observed. Example 2 illustrates that data collection difficulties, confidentiality of the personal information, and grouping during the dataset construction often lead to aggregated variables. This is especially true for the public-used dataset.

The rest of the paper is organized as follows. Section 2 presents the ACD model. Section 3 derives the full-information likelihood function and discusses conditions for separability. Section 4 proposes a competing Bayesian estimator using the Gibbs sampler. Section 5 briefly reviews the traditional least squares based solutions. Section 6 compares various estimators by Monte Carlo experiments. Section 7 extends the model to multiple aggregated covariates, imbalanced aggregation, as well as partial aggregation. Section 8 concludes the paper.

2. The ACD model

We follow the model and notation in Hsiao (1979) and Palm and Nijman (1982), but add a richer set of regressors to allow separability of the likelihood function.

The ACD model consists of the following equations:

$$y_{t,i} = x_{t,i}\beta + \mathbf{w}_{t,i}\boldsymbol{\delta} + u_{t,i} \tag{1}$$

$$x_{t,i} = \mathbf{z}_{t,i}\boldsymbol{\alpha} + \mathbf{w}_{t,i}\boldsymbol{\gamma} + v_{t,i} \tag{2}$$

$$\overline{x}_t = \sum_{i=1}^{n} x_{t,i} \tag{3}$$

where

$$\begin{pmatrix} u_{t,i} \\ v_{t,i} \end{pmatrix} \sim i.i.d.N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix} \right], t = 1, ..., T; i = 1, ..., n,$$

$\mathbf{w}_{t,i}, \mathbf{z}_{t,i}$ are exogenous explanatory variables (row vectors), uncorrelated with disturbance terms, and parameters $\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}$ are column vectors.

Eq. (1) is the main regression model and β is the parameter of interest. The complete data of $\{y_{t,i}, \mathbf{w}_{t,i}, \mathbf{z}_{t,i}\}$ are observed, but $\{x_{t,i}\}$ are unavailable, with aggregated values $\{\overline{x}_t\}$ being observed. As in Hsiao (1979), the subscript (t, i) originally refers to the i^{th} observation in the year t. That is, semiannual (or quarterly, monthly) data are aggregated into annual data. In more general settings, we may interpret t as the group index, and i as the i^{th} member in that group. Data of individual members in a group are aggregated. For instance, in Example 2, $x_{t,i}$ refers to the latent pet population in the county i of the state t, but only the state-level population $\overline{x}_t$ is observed.

Eq. (2) is the regression imputation model for the unobserved $x_{t,i}$. The choice of variables for imputation was discussed in Schafer (1997) and Van Buuren et al. (1999). They suggested that covariates in the main regression (i.e. $\mathbf{w}_{t,i}$) should be included, and factors related to missing mechanisms and with substantial explanatory power over $x_{t,i}$ can also be included, which are captured in $\mathbf{z}_{t,i}$. For instance, in Example 1, in addition to $x_{t,i}$ (the latent grant usage) which explains the monthly energy bill, a plausible set of regressors in $\mathbf{w}_{t,i}$ may include the outdoor temperature, household income, family and room size, the age indicator variable, etc. To impute $x_{t,i}$, we may add all variables in $\mathbf{w}_{t,i}$ and monthly saving-to-income ratio as $\mathbf{z}_{t,i}$.

Of course, the data aggregation model *per se* does not require the appearance of $\mathbf{w}_{t,i}$ in both Eq. (1) and (2). Even if some or none of the variables in $\mathbf{w}_{t,i}$ are included in Eq. (2), the model is still estimable by both ML and Bayesian methods. However, the separability of the likelihood and closed-form ML estimator requires the presence of $\mathbf{w}_{t,i}$ in Eq. (2).

The relationship of disturbance terms across two equations determines the role of $x_{t,i}$ in Eq. (1). For a given (t, i), if $u_{t,i}$ and $v_{t,i}$ are correlated, then $x_{t,i}$ is an endogenous regressor in Eq. (1) since $x_{t,i}$ and $u_{t,i}$ are also correlated. Of course, the data aggregation model *per se* is not necessarily associated with endogenity. Maybe $\mathbf{z}_{t,i}$ is included solely because it can better explain and impute the missing $x_{t,i}$. Nevertheless, note that $\mathbf{z}_{t,i}$ satisfies all the requirements of a valid instrument. As we will see below, if we do not restrict $\sigma_{uv} = 0$ and let the data speak for themselves, we allow the separability of the likelihood function.

Though disturbances across equations are allowed to be correlated, through-

out this paper we assume no serial correlation of disturbances. If we had long time series aggregated at varied frequencies, it would be more appropriate to infer the dependency structure of disaggregated series from observed aggregated series. However, in microeconomic applications, the aggregation is often at the geographic or individual level as in Example 2 and 3, and the dependency structure is not obvious. Example 1 does involve temporal aggregation, but there are only 4 or 5 months in winter, so it is harder to model their dependency.

3. Maximum likelihood estimation

3.1. Joint likelihood

Although the ACD model can be estimated by LS procedures, this approach is not efficient (see Palm and Nijman (1982) and Gourieroux and Monfort (1981) for discussion). An efficient estimator of $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha},\beta,\boldsymbol{\gamma},\boldsymbol{\delta},\sigma_u^2,\sigma_v^2,\sigma_{uv})$ can be obtained by making full use of the information conveyed by the observed data, maximizing the joint likelihood

$$\ln L(\boldsymbol{\theta}) = \sum_{t=1}^{T} \ln f(y_{t,1},...,y_{t,n},\overline{x}_t)$$

conditional on the exogenous regressors $\mathbf{w}_{t,i}, \mathbf{z}_{t,i}$.

Hsiao (1979) and Palm and Nijman (1982) derived the likelihood for the case $n = 2$: $\sum_{t=1}^{T} \ln f(y_{t,1}, y_{t,2}, \overline{x}_t)$. Hsiao (1979) first introduced $x_{t,1}$ into the likelihood and then integrated it out: $f(y_{t,1}, y_{t,2}, \overline{x}_t) = \int f(y_{t,1}, y_{t,2} | \overline{x}_t, x_{t,1}) \cdot f(\overline{x}_t, x_{t,1})\, dx_{t,1}$. Palm and Nijman (1982) derived an equivalent form of the likelihood $f(y_{t,1} + y_{t,2}, y_{t,1} - y_{t,2}, \overline{x}_t)$ by integration with respect to $x_{t,1}$.

In fact, a shortcut to obtain the joint likelihood is by manipulation of Eq. (1) to (3).

First of all, define the following symbols:

$$\begin{aligned}
\overline{y}_t &= \sum_{i=1}^{n} y_{t,i},\ \overline{\mathbf{z}}_t = \sum_{i=1}^{n} \mathbf{z}_{t,i},\ \overline{\mathbf{w}}_t = \sum_{i=1}^{n} \mathbf{w}_{t,i}, \\
\mathbf{y}_t &= \begin{pmatrix} y_{t,1} \\ ... \\ y_{t,n} \end{pmatrix},\ \mathbf{x}_t = \begin{pmatrix} x_{t,1} \\ ... \\ x_{t,n} \end{pmatrix},\ \mathbf{z}_t = \begin{pmatrix} \mathbf{z}_{t,1} \\ ... \\ \mathbf{z}_{t,n} \end{pmatrix},\ \mathbf{w}_t = \begin{pmatrix} \mathbf{w}_{t,1} \\ ... \\ \mathbf{w}_{t,n} \end{pmatrix}.
\end{aligned}$$

Plugging Eq. (2) into Eq. (1) and (3), we have

$$\begin{aligned}
y_{t,i} &= \mathbf{z}_{t,i}\boldsymbol{\alpha}\beta + \mathbf{w}_{t,i}\left(\beta\boldsymbol{\gamma} + \boldsymbol{\delta}\right) + \left(\beta v_{t,i} + u_{t,i}\right), \\
\overline{x}_t &= \overline{\mathbf{z}}_t\boldsymbol{\alpha} + \overline{\mathbf{w}}_\mathbf{t}\boldsymbol{\gamma} + \left(v_{t,1} + ... + v_{t,n}\right).
\end{aligned}$$

Since $(v_{t,1}, ..., v_{t,n}, u_{t,1}, ..., u_{t,n})$ can be viewed as $2n$ dimensional multivariate normal, their $n+1$ dimensional (mean-adjusted) linear combinations $(y_{t,1}, ..., y_{t,n}, \overline{x}_t)$ are also multivariate normal, and we have

$$\begin{pmatrix} \mathbf{y}_t \\ \overline{x}_t \end{pmatrix} \sim N\left\{ \begin{bmatrix} \mathbf{z}_t\boldsymbol{\alpha}\beta + \mathbf{w}_t\left(\beta\boldsymbol{\gamma} + \boldsymbol{\delta}\right) \\ \overline{\mathbf{z}}_t\boldsymbol{\alpha} + \overline{\mathbf{w}}_\mathbf{t}\boldsymbol{\gamma} \end{bmatrix}, \begin{bmatrix} \left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)\mathbf{I}_n & \left(\beta\sigma_v^2 + \sigma_{uv}\right)\boldsymbol{\iota}_n \\ \left(\beta\sigma_v^2 + \sigma_{uv}\right)\boldsymbol{\iota}_n' & n\sigma_v^2 \end{bmatrix} \right\},$$

where $\mathbf{I}_n$ is the identity matrix, and $\boldsymbol{\iota}_n$ is a column vector of ones.

If we decompose the joint multivariate normal density into

$$f\left(\mathbf{y}_t, \overline{x}_t\right) = f\left(\mathbf{y}_t \left| \overline{x}_t \right.\right) \cdot f\left(\overline{x}_t\right),$$

we will arrive at expression (11) on p.246 in Hsiao (1979) where $f(\mathbf{y}_t|\overline{x}_t)$ is termed L_1 and $f(\overline{x}_t)$ termed L_2. Palm and Nijman (1982) had the same in expression (4) on p.335.

3.2. Separability of likelihood

The likelihood function $L(\boldsymbol{\theta})$ is separable if it can be factorized as

$$L(\boldsymbol{\theta}) = L_1(\boldsymbol{\theta}_1) \cdot L_2(\boldsymbol{\theta}_2),$$

where $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is a partition of $\boldsymbol{\theta}$.

A separable likelihood function has a computational advantage in that maximization with respect to $\boldsymbol{\theta}$ can be performed through $\max_{\boldsymbol{\theta}_1} L_1(\boldsymbol{\theta}_1)$ and $\max_{\boldsymbol{\theta}_2} L_2(\boldsymbol{\theta}_2)$ respectively. Moreover, Anderson (1957) discovered that those two maximizations may have analytic solutions for some (but not all) types of missing multivariate normal variates.

For the ACD model, Palm and Nijman (1982) pointed out that the L_1 and L_2 in Hsiao (1979) are not separable. However, there are two useful special cases in which separability does exist. To find the separable form, we first factorize the joint density in the other order:

$$f(\mathbf{y}_t, \overline{x}_t) = f(\mathbf{y}_t) \cdot f(\overline{x}_t|\mathbf{y}_t),$$

and then we reparameterize the model and construct the partition.

The first case is when $z_{t,i}$ is a scalar variable (so is α), and no restrictions are imposed on σ_{uv}.

Then we have

$$f(\mathbf{y}_t) = \phi(\mathbf{y}_t; \mathbf{z}_t \cdot A + \mathbf{w}_t \cdot \mathbf{B}, C \cdot \mathbf{I}_n),$$
$$f(\overline{x}_t | \mathbf{y}_t) = \phi(\overline{x}_t; \overline{z}_t \cdot D + \overline{\mathbf{w}}_t \cdot \mathbf{E} + \overline{y}_t \cdot F, G),$$

where $\phi(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ evaluated at $\mathbf{y}$, and

$$\begin{aligned}
A &= \alpha\beta, \\
\mathbf{B} &= \beta\boldsymbol{\gamma} + \boldsymbol{\delta}, \\
C &= \beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}, \\
D &= \alpha - \left(\beta\sigma_v^2 + \sigma_{uv}\right)\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}\alpha\beta, \\
\mathbf{E} &= \boldsymbol{\gamma} - \left(\beta\sigma_v^2 + \sigma_{uv}\right)\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}(\beta\boldsymbol{\gamma} + \boldsymbol{\delta}), \\
F &= \left(\beta\sigma_v^2 + \sigma_{uv}\right)\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}, \\
G &= n\sigma_v^2 - n\left(\beta\sigma_v^2 + \sigma_{uv}\right)^2\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}.
\end{aligned}$$

The derivation is straightforward in that $f(\mathbf{y}_t)$ and $f(\overline{x}_t | \mathbf{y}_t)$ are simply the marginal and conditional density of the multivariate normal distribution. However, the result implies that the likelihood function have a separable form with respect to the new parameters, which can be partitioned as $(A, \mathbf{B}, C), (D, \mathbf{E}, F, G)$.

Furthermore, note that

$$\max_{A,\mathbf{B},C} \sum_{t=1}^{T} \ln f(\mathbf{y}_t)$$

is equivalent to the ML estimation of the linear regression

$$y_{t,i} = z_{t,i} \cdot A + \mathbf{w}_{t,i} \cdot \mathbf{B} + u_{t,i}.$$

The analytic ML estimator is given by

$$\begin{pmatrix} \widehat{A} \\ \widehat{\mathbf{B}} \end{pmatrix} = \left[\sum_{t=1}^{T} \sum_{i=1}^{n} \left(z_{t,i}, \mathbf{w}_{t,i} \right)' \left(z_{t,i}, \mathbf{w}_{t,i} \right) \right]^{-1} \left[\sum_{t=1}^{T} \sum_{i=1}^{n} \left(z_{t,i}, \mathbf{w}_{t,i} \right)' y_{t,i} \right],$$

$$\widehat{C} = \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \left(y_{t,i} - z_{t,i} \widehat{A} - \mathbf{w}_{t,i} \widehat{\mathbf{B}} \right)^2.$$

Similarly,

$$\max_{D,\mathbf{E},F,G} \sum_{t=1}^{T} \ln f \left(\overline{x}_t \left| \mathbf{y}_t \right. \right)$$

is equivalent to the ML estimation of the linear regression

$$\overline{x}_t = \overline{z}_t \cdot D + \overline{\mathbf{w}}_t \cdot \mathbf{E} + \overline{y}_t \cdot F + \varepsilon_t,$$

with the estimator given by

$$\begin{pmatrix} \widehat{D} \\ \widehat{\mathbf{E}} \\ \widehat{F} \end{pmatrix} = \left[\sum_{t=1}^{T} \left(\overline{z}_t, \overline{\mathbf{w}}_t, \overline{y}_t \right)' \left(\overline{z}_t, \overline{\mathbf{w}}_t, \overline{y}_t \right) \right]^{-1} \left[\sum_{t=1}^{T} \left(\overline{z}_t, \overline{\mathbf{w}}_t, \overline{y}_t \right)' \overline{x}_t \right],$$

$$\widehat{G} = \frac{1}{T} \sum_{t=1}^{T} \left(\overline{x}_t - \overline{z}_t \widehat{D} - \overline{\mathbf{w}}_t \widehat{\mathbf{E}} - \overline{y}_t \widehat{F} \right)^2.$$

Finally, since the ML estimator is invariant to the one-to-one reparameterization, the full-information ML estimator for $(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}, \boldsymbol{\delta}, \sigma_u^2, \sigma_v^2, \sigma_{uv})$ can be solved accordingly. The explicit solution is

$$
\begin{aligned}
\boldsymbol{\alpha} &= D + AF, \\
\beta &= \frac{A}{D + AF}, \\
\boldsymbol{\gamma} &= \mathbf{E} + \mathbf{B}F, \\
\boldsymbol{\delta} &= \frac{\mathbf{B}D - A\mathbf{E}}{D + AF}, \\
\sigma_u^2 &= \frac{A^2 G + nCD^2}{n\left(D + AF\right)^2}, \\
\sigma_v^2 &= CF^2 + \frac{1}{n}G, \\
\sigma_{uv} &= \frac{nCDF - AG}{n\left(D + AF\right)}.
\end{aligned}
$$

Plugging in the estimated values $\left(\widehat{A}, \widehat{\mathbf{B}}, \widehat{C}, \widehat{D}, \widehat{\mathbf{E}}, \widehat{F}, \widehat{G}\right)$, we will obtain the ML estimator $\left(\widehat{\boldsymbol{\alpha}}, \widehat{\beta}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\delta}}, \widehat{\sigma_u^2}, \widehat{\sigma_v^2}, \widehat{\sigma_{uv}}\right)$.

There is one issue we need to clarify. Note that the covariance matrix of $u_{t,i}, v_{t,i}$ must be positive definite, which imposes inequality constraints $\sigma_u^2 > 0$, $\sigma_v^2 > 0$, and $\sigma_u^2\sigma_v^2 - \sigma_{uv}^2 > 0$. By rewriting $\widehat{\sigma_u^2}$, $\widehat{\sigma_v^2}$, $\widehat{\sigma_u^2\sigma_v^2} - \widehat{\sigma_{uv}}$ in terms of $\left(\widehat{A}, \widehat{\mathbf{B}}, \widehat{C}, \widehat{D}, \widehat{\mathbf{E}}, \widehat{F}, \widehat{G}\right)$, we can verify the three expressions are always strictly positive. In fact, the marginal distribution $f\left(\mathbf{y}_t\right)$ and the conditional distribution $f\left(\overline{x}_t \left| \mathbf{y}_t \right.\right)$ are mechanically estimated by LS with $\widehat{C} > 0$, $\widehat{G} > 0$ by construction. Therefore, the joint distribution $f\left(\mathbf{y}_t, \overline{x}_t\right)$ is always well defined with positive definite covariance matrix. Furthermore, a little algebra reveals the positive definite covariance matrix of $\mathbf{y}_t, \overline{x}_t$ implies $\sigma_u^2\sigma_v^2 - \sigma_{uv}^2 > 0$. In a word, the above procedure guarantees that inequality constraints are automatically satisfied.

The second case is that $z_{t,i}$ does not exist, and σ_{uv} is restricted to zero.

Then we have

$$f(\mathbf{y}_t) = \phi(\mathbf{y}_t; \mathbf{w}_t \cdot \mathbf{B}, C \cdot \mathbf{I}_n),$$
$$f(\overline{x}_t | \mathbf{y}_t) = \phi(\overline{x}_t; \overline{\mathbf{w}}_t \cdot \mathbf{E} + \overline{y}_t \cdot F, G),$$

where

$$\begin{aligned}
\mathbf{B} &= \beta\boldsymbol{\gamma} + \boldsymbol{\delta}, \\
C &= \beta^2\sigma_v^2 + \sigma_u^2, \\
\mathbf{E} &= \boldsymbol{\gamma} - \left(\beta\sigma_v^2\right)\left(\beta^2\sigma_v^2 + \sigma_u^2\right)^{-1}(\beta\boldsymbol{\gamma} + \boldsymbol{\delta}), \\
F &= \left(\beta\sigma_v^2\right)\left(\beta^2\sigma_v^2 + \sigma_u^2\right)^{-1}, \\
G &= n\sigma_v^2 - n\left(\beta\sigma_v^2\right)^2\left(\beta^2\sigma_v^2 + \sigma_u^2\right)^{-1}.
\end{aligned}$$

The separability of the likelihood implies $\widehat{\mathbf{B}}, \widehat{C}$ coming from the linear regression $y_{t,i} = \mathbf{w}_{t,i} \cdot \mathbf{B} + u_{t,i}$, and $\widehat{\mathbf{E}}, \widehat{F}, \widehat{G}$ coming from $\overline{x}_t = \overline{\mathbf{w}}_t \cdot \mathbf{E} + \overline{y}_t \cdot F + \varepsilon_t$. The full-information ML estimator of $(\beta, \boldsymbol{\gamma}, \boldsymbol{\delta}, \sigma_u^2, \sigma_v^2)$ can be solved with the following closed form:

$$\begin{aligned}
\beta &= \frac{nCF}{nCF^2 + G}, \\
\boldsymbol{\gamma} &= \mathbf{E} + \mathbf{B}F, \\
\boldsymbol{\delta} &= \frac{\mathbf{B}G - nC\mathbf{E}F}{nCF^2 + G}, \\
\sigma_u^2 &= \frac{CG}{nCF^2 + G}, \\
\sigma_v^2 &= CF^2 + \frac{1}{n}G.
\end{aligned}$$

The above two cases deserve some remarks.

First, if the ACD model specification does not belong to the two special cases, it does not mean the model is not estimable by ML. As long as the model is identifiable, the likelihood can always be maximized by numerical procedures. However, separability of the likelihood offers a computational advantage — it is even less costly than the imputed value two-step estimator. Note that both point estimators are computed from two OLS regressions, but the standard error adjustment for the two-step estimator is not straightforward, while standard errors for the ML estimator can be computed by the Delta method.

Second, as far as an applied problem is concerned, the two special cases are not so restrictive as they might seem. Separability of the likelihood can be achieved if we reasonably redesign the model in use. Case 1 requires one instrument variable $z_{t,i}$ per each one endogenous aggregated regressor $x_{t,i}$. Suppose the goal is to impute the aggregated $x_{t,i}$, but the endogenity is not of major concern. By allowing the possibility of non-zero σ_{uv}, we gain, in addition to $\mathbf{w}_{t,i}$, another variable $z_{t,i}$ which lends explanatory power to impute $x_{t,i}$. If more than one such additional variables are available, we might consider extracting the first principal component of them. In that fashion we retain most of the explanatory power for imputation and meanwhile save computational costs. Case 2 is suitable when $x_{t,i}$ is not endogenous, but the only set of regressors $\mathbf{w}_{t,i}$ appearing in both Eq. (1) and (2) seems restrictive. If we have different covariates for the two equations in mind, we can take the union of the regressors to form $\mathbf{w}_{t,i}$.

Lastly, if $z_{t,i}$ does not exist, and we allow $\sigma_{uv} \neq 0$, the model is not

identified, and thus should be avoided.

4. Bayesian estimator

If conditions of the either special cases are satisfied, the analytic ML estimator is the first choice of estimating the ACD model for the sake of efficiency and computability. However, we are aware that there are circumstances when i) we cannot revise our model catering to the special cases, and the numerical ML does not perform satisfactorily; ii) we have prior information or beliefs on the parameter values or restrictions on the parameters; or iii) we are primarily estimating other models, meanwhile some data are aggregated. In these situations, the likelihood might be difficult to formulate and maximize numerically. We therefore propose a competing Bayesian approach to handle an aggregated covariate, where the joint posteriors of the latent covariate as well as other parameters of uncertainty are simulated using the Gibbs sampler.

Despite that the frequentist and Bayesian inference treat parameter uncertainty in fundamentally different ways, the structure of information contained in the sampling distribution is the same. If the priors are diffuse, the posteriors are mostly learned from the likelihood. Taking a pragmatic stance, the ML and Bayesian inference is comparable in the sense that the large-sample posterior mean (or mode) should be close to the ML estimator and the posterior variance be close to the inverse of the information matrix of the likelihood function.

The Gibbs sampler cycles through the full conditional posteriors (each variable or variables block conditional on other variables as well as the data).

In latent variable models, posterior conditionals for model parameters would be of standard form if the latent variable were known. The key step is to specify the posterior conditionals for the latent variable.

Let us first define the following symbols:

$$\boldsymbol{\psi} = (\beta, \boldsymbol{\delta}', \boldsymbol{\alpha}', \boldsymbol{\gamma}')', \mathbf{X}_{ti} = \begin{bmatrix} (x_{t,i}, \mathbf{w}_{t,i}) & \mathbf{0} \\ \mathbf{0} & (\mathbf{z}_{t,i}, \mathbf{w}_{t,i}) \end{bmatrix},$$

$$\mathbf{Y}_{t,i} = \begin{pmatrix} y_{t,i} \\ x_{t,i} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}.$$

For illustration, we specify the following proper priors:

$$\boldsymbol{\psi} \sim N\left(\underline{\boldsymbol{\mu}}, \underline{\mathbf{V}}\right),$$
$$\boldsymbol{\Sigma}^{-1} \sim Wishart\left(\underline{\boldsymbol{\Omega}}, \underline{\nu}\right).$$

Conditional on the latent $\{x_{t,i}\}$, it is a standard seemingly unrelated regression (SUR) model. The full posterior conditionals are (refer to the textbook Koop et al. (2007) for a derivation):

$$\boldsymbol{\psi} \,|\cdot \sim N\left(\mathbf{Dd}, \mathbf{D}\right),$$
$$\boldsymbol{\Sigma}^{-1} \,|\cdot \sim Wishart\left(\overline{\boldsymbol{\Omega}}, \overline{\nu}\right),$$

where

$$
\begin{aligned}
\mathbf{D} &= \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{X}_{ti}' \boldsymbol{\Sigma}^{-1} \mathbf{X}_{ti} + \underline{\mathbf{V}}^{-1} \right)^{-1}, \\
\mathbf{d} &= \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{X}_{ti}' \boldsymbol{\Sigma}^{-1} \mathbf{Y}_{t,i} + \underline{\mathbf{V}}^{-1} \underline{\boldsymbol{\mu}}, \\
\overline{\boldsymbol{\Omega}} &= \left[\underline{\boldsymbol{\Omega}}^{-1} + \sum_{t=1}^{T} \sum_{i=1}^{n} \left(\mathbf{Y}_{t,i} - \mathbf{X}_{ti} \boldsymbol{\psi} \right) \left(\mathbf{Y}_{t,i} - \mathbf{X}_{ti} \boldsymbol{\psi} \right)' \right]^{-1}, \\
\overline{\nu} &= \underline{\nu} + nT.
\end{aligned}
$$

To derive the full posterior conditional distribution for the latent $\{x_{t,i}\}$, we first introduce a proposition on restricted multivariate normal distribution. Fraser (1951) solved for the general form of n-dimension distribution subject to k ($k < n$) linear constraints by transforming the linear space, but his procedure is descriptive and no explicit distributional forms are given. However, for the purpose of this paper, we only need to solve for a special case — n originally uncorrelated normal variates subject to one aggregation constraint. Explicit solutions are provided in the following proposition (See the appendix for a proof).

Proposition 1. *Let* $\mathbf{x} = (x_1, ..., x_n)'$ *be a multivariate normal random vector with zero correlations.* $x_i \sim N(\mu_i, \sigma^2), i = 1, ..., n$. *Conditional on the aggregation constraint:* $\sum_{i=1}^{n} x_i = \overline{x}$ *where* $\overline{x}$ *is fixed, we have*

$$
\mathbf{x}_{-n} | \overline{x} \sim N \left[\boldsymbol{\mu}_{-n} + \frac{1}{n} \left(\overline{x} - \sum_{i=1}^{n} \mu_i \right) \boldsymbol{\iota}_{n-1}, \sigma^2 \left(\mathbf{I}_{n-1} - \frac{1}{n} \boldsymbol{\iota}_{n-1} \boldsymbol{\iota}_{n-1}' \right) \right],
$$

where $\mathbf{x}_{-n} = (x_1, ..., x_{n-1})'$, $\boldsymbol{\mu}_{-n} = (\mu_1, ..., \mu_{n-1})'$, $\mathbf{I}_{n-1}$ *is the identity matrix, and* $\boldsymbol{\iota}_{n-1}$ *is a vector of ones. Moreover,* $x_n | \mathbf{x}_{-n}, \overline{x}$ *is degenerated, and equals to* $\overline{x} - \sum_{i=1}^{n-1} x_i$.

Marginally, each $x_i \,|\overline{x}$ is $N\left[\mu_i + \frac{1}{n}\left(\overline{x} - \sum_{i=1}^{n}\mu_i\right), \left(1-\frac{1}{n}\right)\sigma^2\right]$ for $i = 1, ..., n$. However, only $n-1$ of them can form a multivariate normal distribution, with the remaining variable having a degenerated distribution. To sample from that restricted distribution, we first take draws from $f\left(\mathbf{x}_{-n}\,|\overline{x}\right)$, and then subtract $\sum_{i=1}^{n-1} x_i$ from $\overline{x}$ to obtain x_n.

We now derive the posterior conditional distribution of the latent $\{x_{t,i}\}$ in the ACD model. The latent covariate data are uncorrelated unconditionally, but correlated conditional on the aggregation constraint. So for each group $t = 1, ...T$, we sample $(x_{t,1}, ..., x_{t,n})$ using Proposition 1. The distributional form is provided in the next proposition (See the appendix for a proof).

Proposition 2. *For every group t, the full posterior conditional $\mathbf{x}_t\,|\cdot$ can be decomposed as*

$$\begin{aligned}
\mathbf{x}_{t,-n}\,|\cdot &\sim N\left[\overline{\boldsymbol{\mu}}_{t,-n} + \frac{1}{n}\left(\overline{x}_t - \sum_{i=1}^{n}\overline{\mu}_{t,i}\right)\boldsymbol{\iota}_{n-1}, \overline{\sigma}^2\left(\mathbf{I}_{n-1} - \frac{1}{n}\boldsymbol{\iota}_{n-1}\boldsymbol{\iota}_{n-1}'\right)\right], \\
x_{t,n}\,|\cdot, \mathbf{x}_{t,-n} &= \overline{x}_t - \sum_{i=1}^{n-1} x_{t,i},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{x}_{t,-n} &= (x_{t,1}, ..., x_{t,n-1})', \\
\overline{\boldsymbol{\mu}}_{t,-n} &= \left(\overline{\mu}_{t,1}, ..., \overline{\mu}_{t,n-1}\right)', \\
\overline{\mu}_{t,i} &= \mathbf{z}_{t,i}\boldsymbol{\alpha} + \mathbf{w}_{t,i}\boldsymbol{\gamma} + \frac{\beta\sigma_v^2 + \sigma_{uv}}{\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}}\left[y_{t,i} - \mathbf{z}_{t,i}\boldsymbol{\alpha}\beta - \mathbf{w}_{t,i}\left(\beta\boldsymbol{\gamma} + \boldsymbol{\delta}\right)\right], \\
\overline{\sigma}^2 &= \sigma_v^2 - \left(\beta\sigma_v^2 + \sigma_{uv}\right)^2\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}.
\end{aligned}$$

The result is conformable with the "Exercise 14.19 (Missing data, 3)" in Koop et al. (2007), That exercise solves a missing data problem with a univariate regression imputation. The intuition underlying the approach is that

our knowledge of missing data is updated by two pieces of information: one from the main regression equation, while the other from the imputation equation. The ACD proceeds further, since there is a third piece of information from the aggregation constraint.

The Gibbs sampler cycles through $\boldsymbol{\psi} \left| \cdot \right.$, $\boldsymbol{\Sigma}^{-1} \left| \cdot \right.$ and $\mathbf{x}_t \left| \cdot \right.$, $t = 1, ..., T$. Once the chain converges, we obtain posterior draws from the joint distribution of $\boldsymbol{\psi}, \boldsymbol{\Sigma}^{-1}, \{\mathbf{x}_t\}$ conditional on $\{y_{t,i}\}, \{\overline{x}_t\}$.

Note that our prior knowledge can be flexibly incorporated into the Bayesian model. For instance, if we know that parameters must belong to some set, we might use truncated priors; if the parameters are subject to equality or inequality constraints, methods in Geweke (1995) can be employed; if we take an objective Bayesian stance, we might use non-informative priors for $\boldsymbol{\psi}$ and $\boldsymbol{\Sigma}^{-1}$. In all those cases, sampling procedures for model parameters may change accordingly, but the essential step to sample the latent $\{x_{t,i}\}$ remains the same.

There are also circumstances when some other models are of primary interest while some data are aggregated. The Bayesian procedure is flexible enough to handle that complicity. For example, when estimating a Probit model where $\{y_{t,i}\}$ is binary, while at the same time one covariate $\{x_{t,i}\}$ is aggregated, the standard Gibbs sampler for the Probit model can still be used, with the insertion of an additional step outlined earlier to sample the latent covariate.

5. Least squares estimators

For completeness of estimation strategies, we outline several ways to estimate the ACD model on the basis of LS.

The first approach is an all-aggregated-data estimator. Since the parameters of primary interest are β and $\boldsymbol{\delta}$, we effectively ignore the imputation regression Eq. (2), but aggregate $y_{t,i}$ and $\mathbf{w}_{t,i}$ as well to regress

$$\overline{y}_t = \overline{x}_t\beta + \overline{\mathbf{w}}_t\boldsymbol{\delta} + \overline{u}_t.$$

This estimator is consistent as $T \to \infty$, the asymptotic variance is n times larger than what it would be attained by regressing Eq. (1) if complete data were observed.

The second approach is a two-step estimator due to Dagenais (1973), which is used to address the conventional missing data problems. In the first step, we use aggregated data $\{\overline{x}_t, \overline{\mathbf{z}}_t, \overline{\mathbf{w}}_t\}$ to fit Eq. (2), and then use disaggregated data $\left\{\mathbf{z}_{t,i}, \mathbf{w}_{t,i}\right\}$ to predict (impute) $\{x_{t,i}\}$ as

$$\widehat{x}_{t,i} = \mathbf{z}_{t,i}\widehat{\boldsymbol{\alpha}} + \mathbf{w}_{t,i}\widehat{\boldsymbol{\gamma}},$$

where $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}$ is the OLS estimator of regressing $\{\overline{\mathbf{z}}_t, \overline{\mathbf{w}}_t\}$ on $\{\overline{x}_t\}$.

In the second step, with $\widehat{x}_{t,i}$ in place of $x_{t,i}$, we regress Eq. (1). Provided that the set $\mathbf{z}_{t,i}$ is non-empty, the Dagenais estimator is consistent. Otherwise, we have perfect multicollinearity and β is not identified. This is a difference between data aggregation and general missing data problems.

There is one obvious problem with this estimator — the imputed $\{x_{t,i}\}$ cannot sum up to $\{\overline{x}_t\}$. Therefore, the information content in the aggregated data is not fully explored.

The third is the minimum mean squared error (MSE) two-step estimator proposed by Hsiao (1979). The estimator is similar to the Dagenais estimator except that the imputed value is given by

$$\widehat{x}_{t,i} = \mathbf{z}_{t,i}\widehat{\boldsymbol{\alpha}} + \mathbf{w}_{t,i}\widehat{\boldsymbol{\gamma}} + \frac{1}{n}\left(\overline{x}_t - \sum_{i=1}^{n} \mathbf{z}_{t,i}\widehat{\boldsymbol{\alpha}} + \mathbf{w}_{t,i}\widehat{\boldsymbol{\gamma}}\right).$$

Essentially, we spread the imputation discrepancy $\overline{x}_t - \sum_{i=1}^{n} \mathbf{z}_{t,i}\widehat{\boldsymbol{\alpha}} + \mathbf{w}_{t,i}\widehat{\boldsymbol{\gamma}}$ evenly across the fitted value $\mathbf{z}_{t,i}\widehat{\boldsymbol{\alpha}} + \mathbf{w}_{t,i}\widehat{\boldsymbol{\gamma}}$. By construction, the aggregation constraint is always satisfied. The rationale of the imputation can be seen in Proposition 1. The imputed value is the conditional mean of $x_{t,i} \,|\overline{x}_t$, hence the minimum MSE. Furthermore the covariance structure of $\mathbf{x}_{t,-n} \,|\overline{x}_t$ implies the negative correlation of imputation errors. Therefore, Hsiao (1979) proposed using GLS in the second step regression. When $\sigma_{uv} = 0$, the covariance matrix of the disturbances is block diagonal, with the covariance within a group (block) given by $\beta^2\sigma_v^2\left(\mathbf{I}_n - \frac{1}{n}\boldsymbol{\iota}_n\boldsymbol{\iota}_n'\right) + \sigma_u^2\mathbf{I}_n$.

Although it seems that the minimum MSE estimator makes the best use of the information and should outperform the other two estimators, in fact that none of the three LS estimators dominates the others. First, if $\sigma_{uv} \neq 0$, the minimum MSE estimator is inconsistent due to endogeneity, while the Dagenais estimator is still consistent. Second, if imputation is of poor quality — σ_v^2 is large, it is possible that Dagenais estimator is less efficient than the all-aggregated-data estimator. We will demonstrate those claims in the appendix.

Lastly, since both the Dagenais estimator and the minimum MSE estimator replace the true $x_{t,i}$ with the imputed value $\widehat{x}_{t,i}$ in Eq. (1), the conventional OLS standard error underestimates the true variability of the

estimator. One solution is to analytically derive a modified standard error by accounting for all uncertainties, and an alternative strategy is to use multiple imputation. In the latter case, we sample $\left(\widehat{\boldsymbol{\alpha}}^*, \widehat{\boldsymbol{\gamma}}^*, \widehat{\sigma}_v^{2*}\right)$ from the distribution of $\left(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}, \widehat{\sigma}_v^2\right)$, and then generate the noise term $v_{t,i}^*$ from $N\left(0, \widehat{\sigma}_v^{2*}\right)$. Therefore, one set of simulated "complete data" for the Dagenais estimator is constructed as

$$\widehat{x}_{t,i}^* = \mathbf{z}_{t,i}\widehat{\boldsymbol{\alpha}}^* + \mathbf{w}_{t,i}\widehat{\boldsymbol{\gamma}}^* + v_{t,i}^*.$$

Similarly, the simulated "complete data" for the minimum MSE estimator is

$$\widehat{x}_{t,i}^* = \mathbf{z}_{t,i}\widehat{\boldsymbol{\alpha}}^* + \mathbf{w}_{t,i}\widehat{\boldsymbol{\gamma}}^* + v_{t,i}^* + \frac{1}{n}\left(\overline{x}_t - \sum_{i=1}^{n} \mathbf{z}_{t,i}\widehat{\boldsymbol{\alpha}}^* + \mathbf{w}_{t,i}\widehat{\boldsymbol{\gamma}}^* + v_{t,i}^*\right).$$

Repeat the process several times, hence several copies of the "complete data". For each copy, estimate Eq. (1) by OLS. The final point estimator is the average of repeated estimates, with the total variance equal to the variance of repeated estimates (the between variability), plus the average of the estimated variances (the within variability).

6. Simulation studies

In this section, we use simulated data to evaluate the performance of various estimators listed in previous sections. For the ACD model with likelihood separability, we compare the analytic ML estimator to three LS estimators, focusing on their relative efficiency. For the model without separability, we compare the performance of the numerical ML and the Gibbs sampler, with the focus on the estimator stability.

For the case with separability, the simulated data experiment is specified as follows:

$n = 12$, $T = 300$,

$y_{t,i} = (x_{t,i}, \mathbf{w}_{t,i}) \cdot (1, 2, 3, 4)' + u_{t,i}$,

$x_{t,i} = (z_{t,i}, \mathbf{w}_{t,i}) \cdot \left(\frac{1}{2}, 1, 1, 1\right)' + v_{t,i}$,

$$\begin{pmatrix} u_{t,i} \\ v_{t,i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.1 \end{pmatrix} \right].$$

$z_{t,i}$ and three components of $\mathbf{w}_{t,i}$ are generated from i.i.d. $N\left(0, \frac{1}{4}\right)$.

For each set of simulated data, we obtain the three LS estimators (i.e. the all-aggregated-data, Dagenais, and minimum MSE estimators) as well as the analytic ML estimator. Then we repeat the data generating process 500 times, hence 500 copies of estimators.

Summary statistics are reported in Table 1. Each estimation approach takes three columns. The first column reports the estimator corresponding to the first simulated data set. The second column shows the average point estimators in 500 repetitions. The third column lists the standard deviation of the 500 repetitions, which can be viewed as the Monte Carlo standard error of the point estimator. If divided by $\sqrt{500}$, it indicates the numerical standard error (NSE) of the average estimator.

In the current setting, $\sigma_{uv} \neq 0$, the minimum MSE estimator is inconsistent (see Section 5). The point estimator $\widehat{\beta}$ averages 1.115 with the NSE 0.0038, significantly different from the true value of 1. Similarly, $\widehat{\boldsymbol{\delta}}$ is also biased due to the endogeneity of $x_{t,i}$.

The simulation results also confirm that both the Dagenais estimator and ML estimator are consistent. $\widehat{\beta}$ using the Dagenais imputation averages

0.998 with the NSE 0.0036, and ML has an average value of 0.996 and NSE 0.0031. Both are close to the true value. However, the Dagenais estimator neglects the aggregation constraint and information usage is inadequate, so we observe that the Monte Carlo standard error of Dagenais estimator is 0.081, which is larger than that of the ML estimator which is 0.069.

The presence of the correlation between disturbances across equations biases the minimum MSE estimator and all-aggregated-data OLS estimator. However, both are consistent when $\sigma_{uv} = 0$. Results when σ_{uv} is changed from 0.1 to 0 are shown in Table 2. On average, $\widehat{\beta}$ for the all-aggregated-data, Dagenais, and minimum MSE estimators are 1.001, 0.997, 0.997 respectively. However, the standard errors are 0.105, 0.084, 0.076 respectively. The minimum MSE estimator incorporates the information content of both $z_{t,i}\widehat{\alpha}$ and $\overline{x}_t$, and therefore outperforms the other two. Also note that the likelihood is not separable when $\sigma_{uv} = 0$, if we still use analytic ML we have to estimate σ_{uv} as well, which is a source of efficiency loss.

		OLS			Dagenais			Minimum MSE			ML		
	True	1st	mean	std	1st	mean	std	1st	mean	std	1st	mean	std
β	1	1.576	1.614	0.099	1.008	0.998	0.081	1.133	1.115	0.085	0.983	0.996	0.069
δ_1	2	1.425	1.386	0.123	1.992	1.999	0.091	1.866	1.882	0.095	2.019	2.003	0.077
δ_2	3	2.312	2.391	0.124	2.968	3.004	0.091	2.842	2.887	0.096	2.970	3.007	0.076
δ_3	4	3.421	3.392	0.130	3.921	4.004	0.091	3.786	3.887	0.096	3.985	4.007	0.078
σ_u^2	0.5	0.412	0.434	0.036	0.412	0.434	0.036	0.412	0.434	0.036	0.497	0.500	0.026
α_1	0.5				0.483	0.502	0.036	0.483	0.502	0.036	0.495	0.502	0.028
γ_1	1				1.011	1.003	0.036	1.011	1.003	0.036	1.010	1.002	0.027
γ_2	1				0.993	1.000	0.037	0.993	1.000	0.037	1.017	0.999	0.027
γ_3	1				1.075	0.998	0.038	1.075	0.998	0.038	1.037	0.997	0.027
σ_v^2	0.1				0.105	0.099	0.008	0.105	0.099	0.008	0.108	0.100	0.007
σ_{uv}	0.1				0.153	0.135	0.026	0.124	0.110	0.024	0.115	0.101	0.010

Table 1: Monte Carlo comparsion of LS and ML estimators, $\sigma_{uv} \neq 0$

The results are based on 500 simulations. Each estimation approach takes three columns. The first column reports the estimator for the first simulated data set. The second and third column show the average and standard deviation of the 500 repetitions. For the aggregated data OLS estimator, only Eq. (1) is estimated. For the Dagenais and minimum MSE estimator, Eq. (2) is regressed identically, so the numbers are the same. The estimated σ_{uv} is obtained from the identity: $Var\left(\beta v_{t,i} + u_{t,i}\right) = \beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}$, where $Var\left(\beta v_{t,i} + u_{t,i}\right)$ is estimated from the regression of $\{y_{t,i}\}$ on $\{z_{t,i}, \mathbf{w}_{t,i}\}$, σ_v^2 is estimated from regressing $\{\overline{x}_t\}$ on $\{\overline{z}_t, \overline{\mathbf{w}}_t\}$, and σ_u^2 from regressing $\{\overline{y}_t\}$ on $\{\overline{x}_t, \overline{\mathbf{w}}_t\}$. In the current setting with $\sigma_{uv} = 0.1$, only the Dagenais and ML estimators are consistent.

		OLS			Dagenais			Minimum MSE			ML		
	True	1st	mean	std	1st	mean	std	1st	mean	std	1st	mean	std
β	1	1.034	1.001	0.105	1.001	0.997	0.084	1.015	0.997	0.076	0.985	0.997	0.081
δ_1	2	1.972	2.001	0.130	1.994	2.000	0.093	1.980	2.000	0.086	2.012	2.002	0.090
δ_2	3	2.853	3.003	0.128	2.957	3.005	0.093	2.943	3.005	0.086	2.960	3.006	0.090
δ_3	4	4.030	4.002	0.137	3.931	4.006	0.093	3.917	4.006	0.087	3.975	4.006	0.091
σ_u^2	0.5	0.468	0.495	0.041	0.468	0.495	0.041	0.468	0.495	0.041	0.486	0.499	0.025
α_1	0.5				0.490	0.503	0.037	0.490	0.503	0.037	0.498	0.503	0.035
γ_1	1				1.011	1.003	0.037	1.011	1.003	0.037	1.009	1.002	0.034
γ_2	1				1.017	0.999	0.037	1.017	0.999	0.037	1.031	0.998	0.034
γ_3	1				1.068	0.997	0.038	1.068	0.997	0.038	1.041	0.997	0.035
σ_v^2	0.1				0.103	0.099	0.008	0.103	0.099	0.008	0.104	0.099	0.008
σ_{uv}	0				0.025	0.004	0.022	0.023	0.004	0.021	0.017	0.002	0.014

Table 2: Monte Carlo comparsion of LS and ML estimators, $\sigma_{uv} = 0$

The results are based on 500 simulations. Each estimation approach takes three columns. The first column reports the estimator for the first simulated data set. The second and third column show the average and standard deviation of the 500 repetitions. For the aggregated data OLS estimator, only Eq. (1) is estimated. For the Dagenais and minimum MSE estimator, Eq. (2) is regressed identically, so the numbers are the same. The estimated σ_{uv} is obtained from the identity: $Var\left(\beta v_{t,i} + u_{t,i}\right) = \beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}$, where $Var\left(\beta v_{t,i} + u_{t,i}\right)$ is estimated from the regression of $\{y_{t,i}\}$ on $\{z_{t,i}, \mathbf{w}_{t,i}\}$, σ_v^2 is estimated from regressing $\{\overline{x}_t\}$ on $\{\overline{z}_t, \overline{\mathbf{w}}_t\}$, and σ_u^2 from regressing $\{\overline{y}_t\}$ on $\{\overline{x}_t, \overline{\mathbf{w}}_t\}$. In the current setting with $\sigma_{uv} = 0$, all estimators are consistent.

When the likelihood is not separable, we compare the performance of the Newton-type numerical ML and Bayesian estimator using the Gibbs sampler. We consider a model without separability by adding one covariate in $\mathbf{z}_{t,i}$. $x_{t,i} = (\mathbf{z}_{t,i}, \mathbf{w}_{t,i}) \cdot (1,1,1,1,1)' + v_{t,i}$. Other settings remain the same.

Though the traditional and Bayesian inference differ fundamentally on the parameter uncertainty, both of the them fully use the sampling information. If the priors are rather diffuse, Bayesian inference should also rely on the full-information likelihood function, and thus in large samples the posterior mean (or mode) should be close to the ML estimator and the posterior standard deviation close to the ML standard error. Here the major concern is to determine which numerical procedure can lead to ideal results in terms of speed and stability.

We specify the prior as $\boldsymbol{\psi} \sim N(\mathbf{0}, 100 \cdot \mathbf{I}_9)$, $\boldsymbol{\Sigma}^{-1} \sim Wishart(\mathbf{I}_2, 1)$, which contains little information compared with the likelihood. The Gibbs sampler is run for 20000 draws with the first half of draws burned in. The convergence and mixing diagnostics reveal that the chain has converged. We treat the posterior mean as the Bayesian point estimator.

We first generate a simulated dataset and set the initial values by adding a $N(0,1)$ disturbance on each true parameter value. If $\boldsymbol{\Sigma}$ is not positive definite, another disturbance draw is taken. Then the generated data and initial values are applied to both ML and the Gibbs sampler. Finally, the process is repeated for 500 times.

As is known, numerical ML can be sensitive to the initial values. In the 500 repetitions of simulated datasets, the numerical ML crashes 14 times and yields another 8 estimates departing far from the true values. Since

		Numeric ML			Bayesian		
	True	1st	mean	std	1st	mean	std
β	1	1.023	1.008	0.060	1.025	1.004	0.024
δ_1	2	1.964	1.993	0.075	1.963	1.998	0.042
δ_2	3	2.997	2.995	0.081	2.997	2.997	0.045
δ_3	4	3.973	3.991	0.077	3.971	3.995	0.041
σ_u^2	0.5	0.493	0.506	0.067	0.495	0.500	0.021
α_1	1	1.013	0.997	0.041	1.013	0.999	0.026
α_2	1	0.999	0.995	0.040	0.998	0.997	0.025
γ_1	1	0.978	1.000	0.030	0.977	0.999	0.025
γ_2	1	0.987	0.999	0.038	0.985	1.000	0.026
γ_3	1	0.968	1.001	0.039	0.967	1.001	0.028
σ_v^2	0.1	0.095	0.101	0.017	0.099	0.104	0.007
σ_{uv}	0.1	0.099	0.096	0.039	0.097	0.098	0.008

Table 3: Monte Carlo comparison of ML and Bayesian estimators

The results are based on 500 simulated data sets. The summary statistics are calculated with the apparent outliers ($\widehat{\beta} < 0$ or $\widehat{\beta} > 2$) removed. Each estimation approach takes three columns. The first column reports the estimator corresponding to the first simulated data set. The second column shows the average of the 500 repetitions. The third column lists the standard deviation of the 500 repetitions.

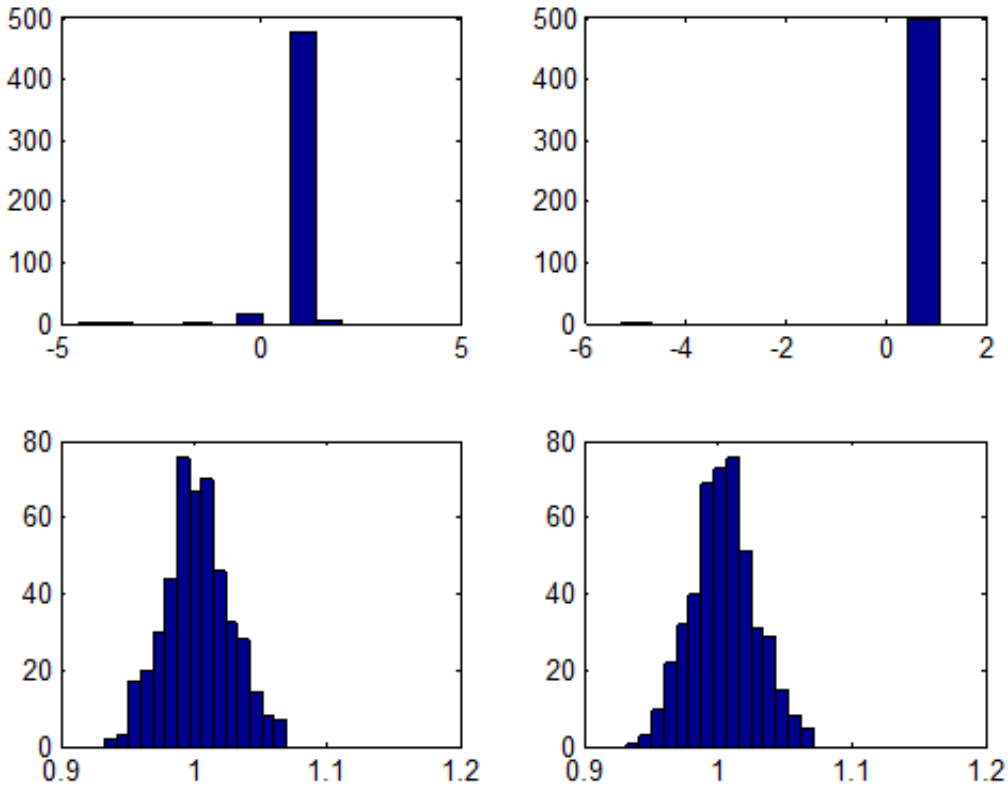

Figure 1: Histogram of $\widehat{\beta}$ in 500 repetitions. The upper left panel shows the ML estimation, and the upper right panel is the Bayesian estimation. The bottom left histogram truncate $\widehat{\beta}$ to (0.93, 1.07) for ML, and bottom right for Bayesian estimator.

the estimator standard deviation is no more than 0.1 and the true value of β is one, we define abnormality to be $\widehat{\beta} \leq 0$ or $\widehat{\beta} \geq 2$. To visualize the departing pattern of abnormal estimators, Figure 1 presents the histogram of $\widehat{\beta}$ in the 500 repetitions. In the case of crash, $\widehat{\beta} = 0$ is assigned for histogram presentation purpose. Compared with the numerical ML, the Gibbs sampler is more stable. It does not crash, and only yields negative $\widehat{\beta}$ twice. The abnormal estimators in the Gibbs sampler are close to each other. It is likely the chain gets stuck in a local high density region and cannot transverse to the region where the true parameters are located.

With the abnormal estimators removed, the summary statistics are presented in Tables 3. The average of the ML and Bayesian estimates are reasonably close to each other. Estimates of $\widehat{\beta}$ average 1.008 for ML and 1.004 for Bayesian. But the standard deviation of the ML estimates is 0.060, larger than that of the Bayesian 0.024. Though the role of the prior distribution and finite draws of the Gibbs sampler may partially explain the smaller variance of the estimator, we do not believe it is the main reason. The numerical issues should be taken into account. With obvious outliers removed, all of the Bayesian estimates lie in (0.93, 1.07) which is about plus/minus 3 standard deviations of the average point estimator. However, we observe several ML estimates with values such as 0.88, 1.22, 1.67. It is not clear it is caused by non-convergence of the optimizer or just caused by the sampling variation. Those values certainly exert a non-negligible impact on the calculation of the sample mean and standard deviation of the estimators. If we truncate the ML $\widehat{\beta}$ in the region (0.93, 1.07) as well, the mean is 1.003 with standard deviation 0.025, which is closer to the inference under the Bayesian scheme.

Although the numerical ML is not as stable as the Gibbs sampler, it does run faster. The average ML estimation time for one set of simulated data is about 1.04 seconds, while the 20000 draws with Gibbs sampler takes an average of 37.9 seconds on an ordinary desktop computer (2.5GHz CPU / 3GB RAM / MATLAB 2009b). Nevertheless, the computation costs for both methods are affordable.

7. Extensions

In an empirical context, the problems we have encountered might be more complicated than the baseline ACD model. In this section, we outline several extensions to the model and ways to handle them.

7.1. Aggregation of several variables

If more than one covariate is aggregated, the model can be extended as

$$\begin{aligned} y_{t,i} &= x_{1,t,i}\beta_1 + ...x_{k,t,i}\beta_k + \mathbf{w}_{t,i}\boldsymbol{\delta} + u_{t,i} \\ x_{1t,i} &= \mathbf{z}_{t,i}\boldsymbol{\alpha}_1 + \mathbf{w}_{t,i}\boldsymbol{\gamma}_1 + v_{1,t,i} \\ &... \\ x_{kt,i} &= \mathbf{z}_{t,i}\boldsymbol{\alpha}_k + \mathbf{w}_{t,i}\boldsymbol{\gamma}_k + v_{k,t,i} \end{aligned}$$

where $\overline{x}_{1,t} = \sum_{i=1}^{n} x_{1,t,i}, \ldots, \overline{x}_{k,t} = \sum_{i=1}^{n} x_{k,t,i}$, $(u_{t,i}, v_{1,t,i}, ..., v_{k,t,i}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.

For ML, we maximize the joint density of observable variables:

$$\ln L(\boldsymbol{\theta}) = \sum_{t=1}^{T} \ln f(y_{t,1}, ..., y_{t,n}, \overline{x}_{1,t}, ..., \overline{x}_{k,t}).$$

The likelihood function is separable in two cases.

The first case is that $\mathbf{z}_{t,i}$ contains exactly k variables and no restrictions are put on $\mathbf{\Sigma}$. The likelihood can be factorized as

$$f\left(\mathbf{y}_t, \overline{x}_{1,t}, ..., \overline{x}_{k,t}\right) = f\left(\mathbf{y}_t\right) \cdot f\left(\overline{x}_{1,t} \left|\mathbf{y}_t\right.\right) \cdot ... \cdot f\left(\overline{x}_{k,t} \left|\mathbf{y}_t, \overline{x}_{1,t}, ..., \overline{x}_{k-1,t}\right.\right).$$

The analytic ML estimator is obtained from the following k regressions:

Regress $y_{t,i}$ on $\mathbf{z}_{t,i}$, $\mathbf{w}_{t,i}$.

Regress $\overline{x}_{1,t}$ on $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$.

Regress $\overline{x}_{2,t}$ on $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$, $\overline{x}_{1,t}$.

... ...

Regress $\overline{x}_{k,t}$ on $\overline{\mathbf{z}}_t$, $\overline{\mathbf{w}}_t$, $\overline{y}_t$, $\overline{x}_{1,t}$,..., $\overline{x}_{k-1,t}$.

The second case is that $\mathbf{z}_{t,i}$ contains J variables ($J < k$), and only J variables may possibly be endogneity regressors in the main regression. In that case, some of the covariance terms in $\mathbf{\Sigma}$ are restricted to zero. To be exact, suppose we believe $x_{1t,i}$ is uncorrelated with $u_{t,i}$, then the first row and first column of $\mathbf{\Sigma}$, except for the diagonal element, are restricted to zero. The reparameterized estimator is obtained from the same regressions as above, but $\mathbf{z}_{t,i}$ and $\overline{\mathbf{z}}_t$ have reduced dimensions.

In words, separability requires that one instrument corresponds to one potentially endogenous variable.

7.2. Unbalanced aggregation

In some applications, group sizes are not equal so that n needs to be written as n_t.

Compared with the ML solutions outlined in Section 3, the separability conditions do not change. The expression of $f(\mathbf{y}_t)$ remains the same, so does the first regression of $y_{t,i}$ on $\mathbf{z}_{t,i}$, $\mathbf{w}_{t,i}$. However, the variance of $\overline{x}_t|\mathbf{y}_t$ changes:

$$f(\overline{x}_t|\mathbf{y}_t) = \phi(\overline{x}_t; \overline{z}_t \cdot D + \overline{\mathbf{w}}_t \cdot \mathbf{E} + \overline{y}_t \cdot F, n_t G),$$

where $n_t G = n_t\sigma_v^2 - n_t(\beta\sigma_v^2 + \sigma_{uv})^2\left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}$. Other components remain the same as before. It implies that

$$\max_{D,\mathbf{E},F,G} \sum_{t=1}^{T} \ln f(\overline{x}_t|\mathbf{y}_t)$$

can be obtained by weighed least squares of $\overline{x}_t$ on $\overline{z}_t, \overline{\mathbf{w}}_t, \overline{y}_t$ with the weights proportional to n_t.

Procedures of the Bayesian simulator are largely unchanged in the unbalanced aggregation. Full posterior conditionals of $\boldsymbol{\psi}$ and $\boldsymbol{\Sigma}^{-1}$ remain the same, and we use n_t instead of n when groupwise taking draws of $\{x_{t,i}\}$.

7.3. Partial aggregation

In applied problems, data aggregation may take on another degree of complexity. For instance, in Example 2, suppose we do find county-level pet population for some states, but not for the rest. How do we make the best use of the incomplete county-level data instead of regressing merely with aggregated state-level data?

In general, the partial aggregation problem is raised as follows: suppose group t has n_t members, among which the first k_t are observable and the rest are missing. In addition, the aggregated value $\overline{x}_t = \sum_{i=1}^{n} x_{t,i}$ is known. The data are generated according to Eq. (1) and (2).

To address this problem, group t can be divided into k_t+1 smaller groups. The first k_t groups are a singleton with known $x_{t,i}$, and the last group contains $n_t - k_t$ members, whose latent values sum up to $\overline{x}_t - \sum_{i=1}^{k_t} x_{t,i}$. Then the problem is equivalent to the unbalanced aggregation introduced in the previous subsection, and both ML and Bayesian estimators can be implemented.

8. Conclusions

Hsiao's model offers a simple framework for addressing data aggregation problems. This paper explores several estimation strategies for this model, showing that the solutions do not always require numerical tools proposed by Palm and Nijman (1982).

The first is a full-information ML estimation. We find that the likelihood function has a separability property in two useful special cases. As long as one instrument corresponds to one endogenous variable, the likelihood can be maximized analytically, with the ML estimator obtained by two linear regressions. For models without the separable likelihood, numerical procedures can also be used, but initial values must be carefully chosen.

The second is the Bayesian simulator implemented by the Gibbs sampler, the advantage of which is two-fold. First, it is more stable. Our Monte Carlo study shows that the Bayesian estimator is less affected by the initial values. Second, it is more flexible. It places no restrictions on the covariates in the imputation regression, and the sampling procedure of latent disaggregated covariates can be easily inserted into researchers' models.

The third is a class of LS estimators. The Dagenais two-step estimator is primarily used for imputing missing data, but it is suitable for aggregated covariate data as well. The minimum MSE estimator is based on the regression imputation, but also uses the aggregation constraint. In the absence of correlation of disturbances among equations, the latter makes better use of information and yields a more precise imputation. Otherwise, the latter is inconsistent, but the former is still consistent. On top of that, the all-aggregated-data OLS method offers the simplest way to estimate the model, and is useful when the imputation is of poor quality. Though conceptually LS is easier to implement than ML and the Gibbs sampler, it is not as efficient in general and thus not recommended unless we cast doubt on the normality assumption.

Appendix A. Proof of Proposition 1

Let $x_i = \mu_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$.

$\overline{x} = \sum_{i=1}^{n} \mu_i + (\varepsilon_1 + ... + \varepsilon_n)$.

Note that $(\varepsilon_1, ..., \varepsilon_n)$ is n dimensional multivariate normal, and so are the n mean-adjusted linear combinations $(x_1, ..., x_{n-1}, \overline{x})$. Then

$$\begin{pmatrix} \mathbf{x}_{-n} \\ \overline{x} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\mu}_{-n} \\ \sum_{i=1}^{n} \mu_i \end{pmatrix}, \begin{pmatrix} \sigma^2 \mathbf{I}_{n-1} & \sigma^2 \boldsymbol{\iota}_{n-1} \\ \sigma^2 \boldsymbol{\iota}'_{n-1} & n\sigma^2 \end{pmatrix} \right].$$

It follows that

$$\mathbf{x}_{-n} | \overline{x} \sim N\left[\boldsymbol{\mu}_{-n} + \frac{1}{n}\left(\overline{x} - \sum_{i=1}^{n} \mu_i \right) \boldsymbol{\iota}_{n-1}, \sigma^2 \left(\mathbf{I}_{n-1} - \frac{1}{n} \boldsymbol{\iota}_{n-1} \boldsymbol{\iota}'_{n-1} \right) \right].$$

Lastly, conditional on $\mathbf{x}_{-n}, \overline{x}$, we have $x_n = \overline{x} - \sum_{i=1}^{n-1} x_i$. □

Appendix B. Proof of Proposition 2

Plugging Eq. (2) into Eq. (1), we have

$$\begin{pmatrix} y_{t,i} \\ x_{t,i} \end{pmatrix} \sim N \left\{ \begin{bmatrix} \mathbf{z}_{t,i}\boldsymbol{\alpha}\beta + \mathbf{w}_{t,i}\left(\beta\boldsymbol{\gamma} + \boldsymbol{\delta}\right) \\ \mathbf{z}_{t,i}\boldsymbol{\alpha} + \mathbf{w}_{t,i}\boldsymbol{\gamma} \end{bmatrix}, \begin{bmatrix} \left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right) & \beta\sigma_v^2 + \sigma_{uv} \\ \beta\sigma_v^2 + \sigma_{uv} & \sigma_v^2 \end{bmatrix} \right\}.$$

It follows that

$$x_{t,i} \left| y_{t,i} \right. \sim N\left(\overline{\mu}_{t,i}, \overline{\sigma}^2\right),$$

where

$$\begin{aligned} \overline{\mu}_{t,i} &= \mathbf{z}_{t,i}\boldsymbol{\alpha} + \mathbf{w}_{t,i}\boldsymbol{\gamma} + \frac{\beta\sigma_v^2 + \sigma_{uv}}{\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}} \left[y_{t,i} - \mathbf{z}_{t,i}\boldsymbol{\alpha}\beta - \mathbf{w}_{t,i}\left(\beta\boldsymbol{\gamma} + \boldsymbol{\delta}\right)\right], \\ \overline{\sigma}^2 &= \sigma_v^2 - \left(\beta\sigma_v^2 + \sigma_{uv}\right)^2 \left(\beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv}\right)^{-1}. \end{aligned}$$

In the presence of the aggregation constraint, we apply Proposition 1. The result follows. □

An alternative proof proceeds by deriving the joint distribution of $\left(\mathbf{y}_t', \mathbf{x}_{t,-n}', \overline{x}_t\right)'$, which is a $2n$ dimensional linear combination of $(v_{t,1}, ..., v_{t,n}, u_{t,1}, ..., u_{t,n})$ and thus still multivariate normal. Therefore, the conditional normal distribution of $\mathbf{x}_{t,-n} \left| \mathbf{y}_t, \overline{x}_t \right.$ can be found after some algebra. The result is the same.

Appendix C. Comparison of least squares estimators

First, we show that if $\sigma_{uv} \neq 0$, the minimum MSE estimator is inconsistent. To see this, we only need to consider the simplest version of the model. Let $n = 2$; α is a known scalar; regressors $\mathbf{w}_{t,i}$ do not exist. The model becomes:

$$\begin{aligned} y_{t,i} &= x_{t,i}\beta + u_{t,i}, \\ x_{t,i} &= z_{t,i}\alpha + v_{t,i}, \\ \overline{x}_t &= x_{t,1} + x_{t,2}. \end{aligned}$$

For the minimum MSE estimator, the imputed value is

$$\begin{aligned} \widehat{x}_{t,i} &= z_{t,i}\alpha + \frac{1}{2}\left[\overline{x}_t - (z_{t,1}\alpha + z_{t,2}\alpha)\right] \\ &= z_{t,i}\alpha + \frac{1}{2}\left(v_{t,1} + v_{t,2}\right). \end{aligned}$$

In the second step, we regress

$$y_{t,i} = \widehat{x}_{t,i}\beta + \varepsilon_{t,i},$$

where $\varepsilon_{t,i} = u_{t,i} + \beta\left[v_{t,i} - \frac{1}{2}\left(v_{t,1} + v_{t,2}\right)\right]$.

The endogenity of $\widehat{x}_{t,i}$ to $\varepsilon_{t,i}$ does not come from the presence of $v_{t,1}, v_{t,2}$ in both expressions, but merely from the correlation between $u_{t,i}$ and $v_{t,i}$. To see this, define $\xi_0 = \frac{1}{2}\left(v_{t,1} + v_{t,2}\right)$, and $\xi_i = v_{t,i} - \frac{1}{2}\left(v_{t,1} + v_{t,2}\right)$, $i = 1, 2$. By change of variable, the joint distribution of ξ_0 and ξ_1 is given by

$$f(\xi_0, \xi_1) = \phi\left(\xi_0 + \xi_1; 0, \sigma_v^2\right) \cdot \phi\left(\xi_0 - \xi_1; 0, \sigma_v^2\right) \cdot |-2|$$
$$\propto \exp\left[-\sigma_v^{-2}\left(\xi_0^2 + \xi_1^2\right)\right].$$

So ξ_0 and ξ_1 are independent and each distributed as $N\left(0, \frac{1}{2}\sigma_v^2\right)$. Similarly, ξ_0 and ξ_2 are independent.

However, as long as $\sigma_{uv} \neq 0$, we have $cov\left(\widehat{x}_{t,i}, \varepsilon_{t,i}\right) = \frac{1}{2}\sigma_{uv}$, hence the endogenous regressor and inconsistent estimator, no matter whether OLS or GLS is used.

In fact, the OLS version of the estimator

$$\widehat{\beta} = \left(\sum_{t=1}^{T}\sum_{i=1}^{n}\widehat{x}_{t,i}^2\right)^{-1}\left(\sum_{t=1}^{T}\sum_{i=1}^{n}\widehat{x}_{t,i}y_{t,i}\right)$$
$$\xrightarrow{p} \beta + \frac{1}{2}\sigma_{uv}\left[\alpha^2 Q_{zz} + \frac{1}{2}\sigma_v^2\right]^{-1},$$

where $Q_{zz} = E\left(z_{t,i}^2\right)$.

On the other hand, the Dagenais estimator is still consistent. The imputed value is $\widetilde{x}_{t,i} = z_{t,i}\alpha$. Then we regress

$$y_{t,i} = \widetilde{x}_{t,i}\beta + \widetilde{\varepsilon}_{t,i},$$

where $\widetilde{\varepsilon}_{t,i} = u_{t,i} + \beta v_{t,i}$, so that $cov\left(\widetilde{x}_{t,i}, \widetilde{\varepsilon}_{t,i}\right) = 0$, even if $\sigma_{uv} \neq 0$.

The estimator

$$\widehat{\beta} = \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \widetilde{x}_{t,i}^2 \right)^{-1} \left(\sum_{t=1}^{T} \sum_{i=1}^{n} \widetilde{x}_{t,i} y_{t,i} \right)$$
$$= \beta + \left[\sum_{t=1}^{T} \sum_{i=1}^{n} \left(z_{t,i}\alpha \right)^2 \right]^{-1} \cdot \left[\sum_{t=1}^{T} \sum_{i=1}^{n} z_{t,i}\alpha \left(u_{t,i} + \beta v_{t,i} \right) \right],$$

so that $\widehat{\beta} \xrightarrow{p} \beta$ and

$$\sqrt{nT} \left(\widehat{\beta} - \beta \right) \xrightarrow{d} N \left[0, \left(\alpha^2 Q_{zz} \right)^{-1} \left(\beta^2 \sigma_v^2 + \sigma_u^2 + 2\beta\sigma_{uv} \right) \right].$$

Clearly, the asymptotic variance of the Dagenais estimator is increasing with σ_v^2. For large enough σ_v^2, it will exceed the variance of all-aggregated-data estimator, which does not use imputation at all. Therefore, if imputation is of poor quality, there is a possibility that the all-aggregated-data estimator is preferred to the Dagenais estimator.

Allison, P. D., 2000. Multiple imputation for missing data: a cautionary tale. Sociological methods and Research 28, 301–309.

Anderson, T. W., 1957. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. Journal of the American Statistical Association 52 (278), 200–203.

Andreou, E., Ghysels, E., Kourtellos, A., 2010. Regression models with mixed sampling frequencies. Journal of Econometrics.

Bureau of Labor Statistics, U.S. Department of Labor, 2010. Occupational Outlook Handbook. Washington: U.S. Government Printing Office.

Dagenais, M. G., 1973. The use of incomplete observations in multiple regression analysis : A generalized least squares approach. Journal of Econometrics 1 (4), 317–328.

Fraser, D. A. S., 1951. Normal samples with linear constraints and given variances. Canadian Journal of Mathematics 3, 363–366.

Geweke, J. F., 1978. Temporal aggregation in the multiple regression model. Econometrica 46 (3), 643–61.

Geweke, J. F., 1995. Bayesian inference for linear models subject to linear inequality constraints. Working Papers 552, Federal Reserve Bank of Minneapolis.

Ghysels, E., Santa-Clara, P., Valkanov, R., 2006. Predicting volatility: getting the most out of return data sampled at different frequencies. Journal of Econometrics 131 (1-2), 59–95.

Gourieroux, C., Monfort, A., 1981. On the problem of missing data in linear models. Review of Economic Studies 48 (4), 579–86.

Hsiao, C., 1979. Linear regression using both temporally aggregated and temporally disaggregated data. Journal of Econometrics 10 (2), 243–252.

Koop, G., Poirier, D. J., Tobias, J. L., 2007. Bayesian Econometric Methods. Cambridge Books. Cambridge University Press.

Palm, F. C., Nijman, T. E., 1982. Linear regression using both temporally aggregated and temporally disaggregated data. Journal of Econometrics 19 (2-3), 333–343.

Rubin, D. B., 1987. Multiple imputation for nonresponse in surveys. New York: Wiley.

Schafer, J. L., 1997. Analysis of incomplete multivariate data. London: Chapman and Hall.

Van Buuren, S., Boshuizen, H. C., Knook, D. L., 1999. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 18, 681–694.